# Assignment 10: Data Scraping

Summer Heschong

## OVERVIEW

This exercise accompanies the lessons in Environmental Data Analytics on data scraping.

## Directions

1. Rename this file `<FirstLast>_A10_DataScraping.Rmd` (replacing `<FirstLast>` with your first and last name).
2. Change "Student Name" on line 3 (above) with your name.
3. Work through the steps, **creating code and output** that fulfill each instruction.
4. Be sure your code is tidy; use line breaks to ensure your code fits in the knitted output.
5. Be sure to **answer the questions** in this assignment document.
6. When you have completed the assignment, **Knit** the text and code into a single PDF file.

## Set up

1. Set up your session:

- Load the packages `tidyverse`, `rvest`, and any others you end up using.
- Check your working directory

```
#1
library(tidyverse)
library(rvest)
library(here)
library(lubridate)

here()
```

```
## [1] "/home/guest/EDE_Fall2024"
```

2. We will be scraping data from the NC DEQs Local Water Supply Planning website, specifically the Durham's 2023 Municipal Local Water Supply Plan (LWSP):

- Navigate to https://www.ncwater.org/WUDC/app/LWSP/search.php
- Scroll down and select the LWSP link next to Durham Municipality.
- Note the web address: https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=03-32-010&year=2023

Indicate this website as the as the URL to be scraped. (In other words, read the contents into an `rvest` webpage object.)

```
#2
website <- read_html(
  'https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=03-32-010&year=2023')
```

3. The data we want to collect are listed below:

- From the "1. System Information" section:
- Water system name
- PWSID
- Ownership
- From the "3. Water Supply Sources" section:
- Maximum Day Use (MGD) - for each month

In the code chunk below scrape these values, assigning them to four separate variables.

> HINT: The first value should be "Durham", the second "03-32-010", the third "Municipality", and the last should be a vector of 12 numeric values (represented as strings)".

```
#3
water_system_name <- website %>%
  html_nodes('div+ table tr:nth-child(1) td:nth-child(2)') %>%
  html_text()
PWSID <- website %>%
  html_nodes('td tr:nth-child(1) td:nth-child(5)') %>%
  html_text()
Ownership <- website %>%
  html_nodes('div+ table tr:nth-child(2) td:nth-child(4)') %>%
  html_text()
Max_day_use <- website %>%
  html_nodes('th~ td+ td') %>%
  html_text()
```

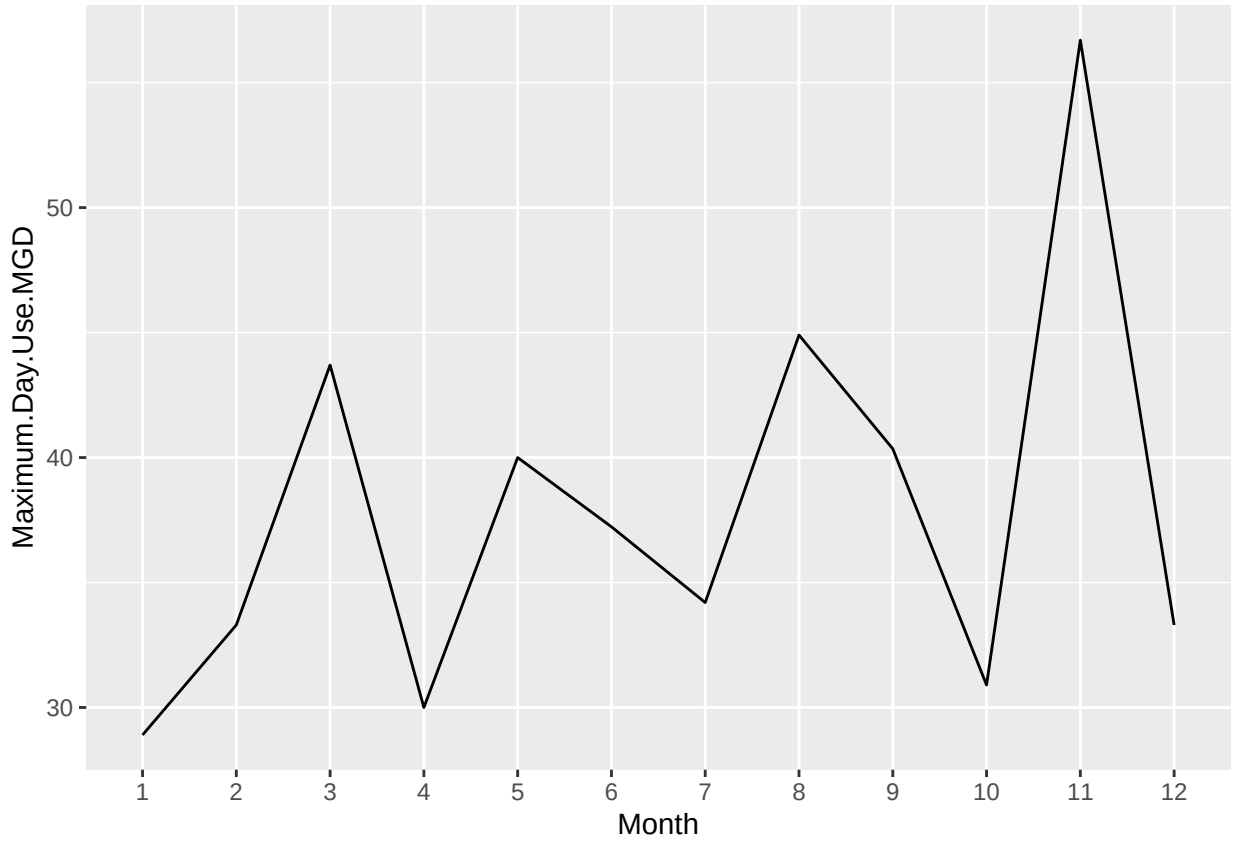
4. Convert your scraped data into a dataframe. This dataframe should have a column for each of the 4 variables scraped and a row for the month corresponding to the withdrawal data. Also add a Date column that includes your month and year in data format. (Feel free to add a Year column too, if you wish.)

   > TIP: Use `rep()` to repeat a value when creating a dataframe.

   > NOTE: It's likely you won't be able to scrape the monthly widthrawal data in chronological order. You can overcome this by creating a month column manually assigning values in the order the data are scraped: "Jan", "May", "Sept", "Feb", etc... Or, you could scrape month values from the web page...

5. Create a line plot of the maximum daily withdrawals across the months for 2023, making sure, the months are presented in proper sequence.

```r
#4 create data frame
water_supply_data <- data.frame(
  'Month' = as.factor(rep(1:12)),
  'Year' = rep(2023, 12),
  'Water System Name' = water_system_name,
  'PWSID' = PWSID,
  'Ownership' = Ownership,
  'Maximum Day Use MGD'= as.numeric(Max_day_use))

#make a column for Date
water_supply_data$Date <- make_date(
    month = water_supply_data$Month, year = water_supply_data$Year)

#change Date to a date object
water_supply_data$Date <- format(water_supply_data$Date, "%Y-%m")

#5 Create line plot

ggplot(water_supply_data, aes(x= Month, y= Maximum.Day.Use.MGD, group = 1)) +
  geom_line()
```

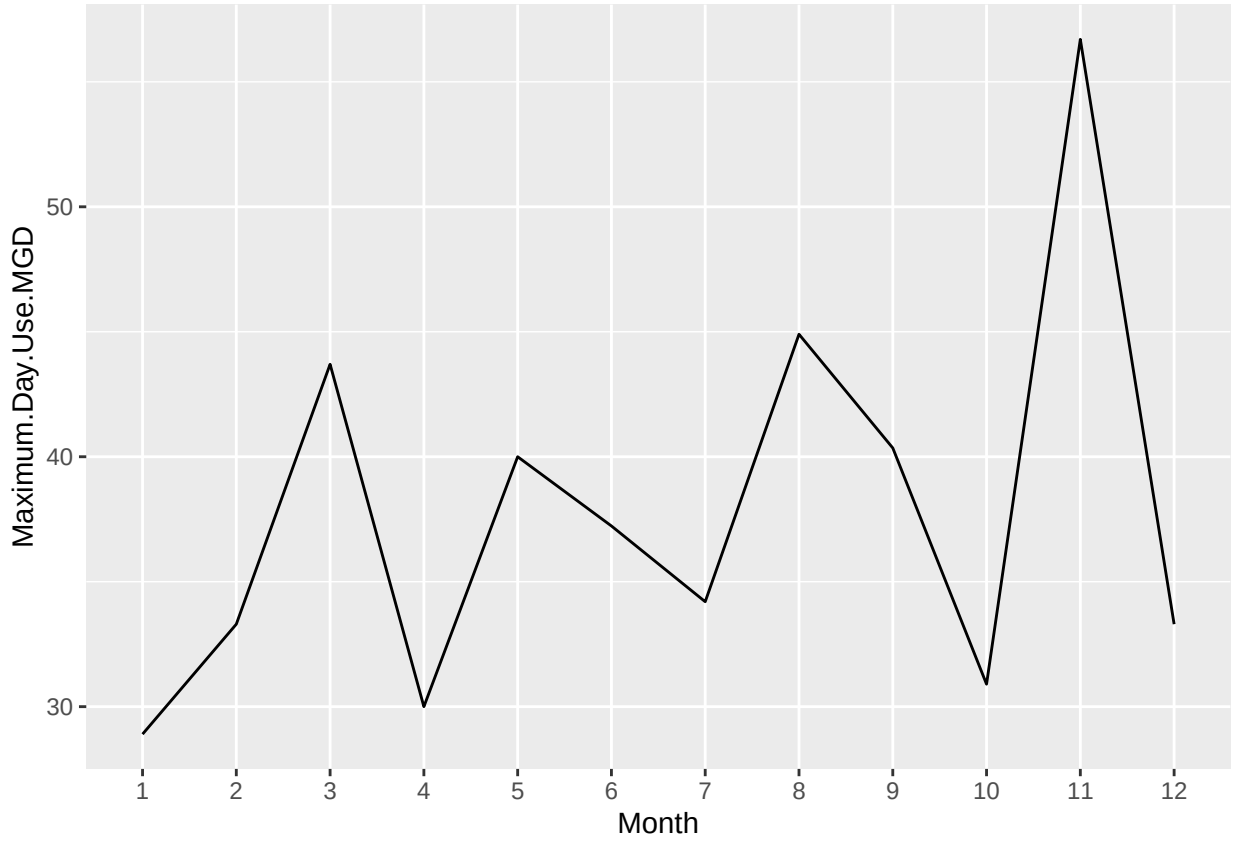



6. Note that the PWSID and the year appear in the web address for the page we scraped. Construct a function using your code above that can scrape data for any PWSID and year for which the NC DEQ has data, returning a dataframe. **Be sure to modify the code to reflect the year and site (pwsid) scraped**.

```r
#6.
scrape_ncwater <- function(the_year, PWSID){

   website <- read_html(
    paste0( 'https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=',PWSID,'&year=',the_year))

  water_system_name_tag <- 'div+ table tr:nth-child(1) td:nth-child(2)'
  PWSID_tag <-'td tr:nth-child(1) td:nth-child(5)'
  Ownership_tag <- 'div+ table tr:nth-child(2) td:nth-child(4)'
  Max_day_use_tag <- 'th~ td+ td'

  water_system_name <- website %>%
  html_nodes('div+ table tr:nth-child(1) td:nth-child(2)') %>%
  html_text()
PWSID <- website %>%
  html_nodes('td tr:nth-child(1) td:nth-child(5)') %>%
  html_text()
Ownership <- website %>%
  html_nodes('div+ table tr:nth-child(2) td:nth-child(4)') %>%
  html_text()
Max_day_use <- website %>%
  html_nodes('th~ td+ td') %>%
  html_text()

water_supply_data <- data.frame(
  'Month' = as.factor(1:12),
  'Year' = rep(the_year, 12),
  'Water System Name' = water_system_name,
  'PWSID' = PWSID,
  'Ownership' = Ownership,
  'Maximum Day Use MGD'= as.numeric(Max_day_use))%>%
  mutate(Date = my(paste(Month,"-",Year)))

Sys.sleep(5)

return(water_supply_data)

}
```

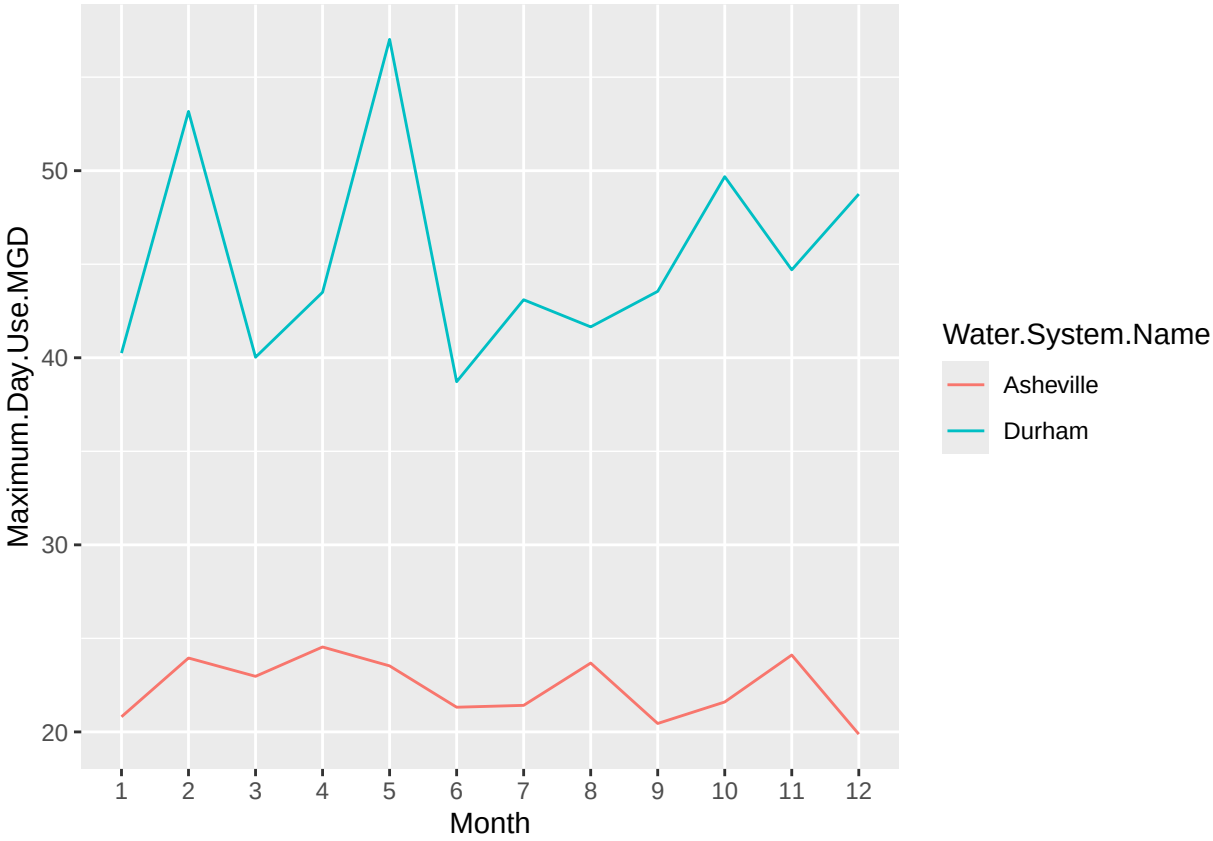
7. Use the function above to extract and plot max daily withdrawals for Durham (PWSID=‘03-32-010’) for each month in 2015

```r
#7
MDW_Durham_2015 <- scrape_ncwater(2015, '03-32-010')

ggplot(water_supply_data, aes(x= Month, y= Maximum.Day.Use.MGD, group = 1)) +
  geom_line()
```




8. Use the function above to extract data for Asheville (PWSID = 01-11-010) in 2015. Combine this data with the Durham data collected above and create a plot that compares Asheville's to Durham's water withdrawals.

```
#8
MDW_Asheville_2015 <- scrape_ncwater(2015, '01-11-010')

MDW_AshandDurham_2015  <- rbind(MDW_Asheville_2015, MDW_Durham_2015)

ggplot(MDW_AshandDurham_2015, aes(x = Month, y= Maximum.Day.Use.MGD, color = Water.System.Name, group= W
  geom_line()
```

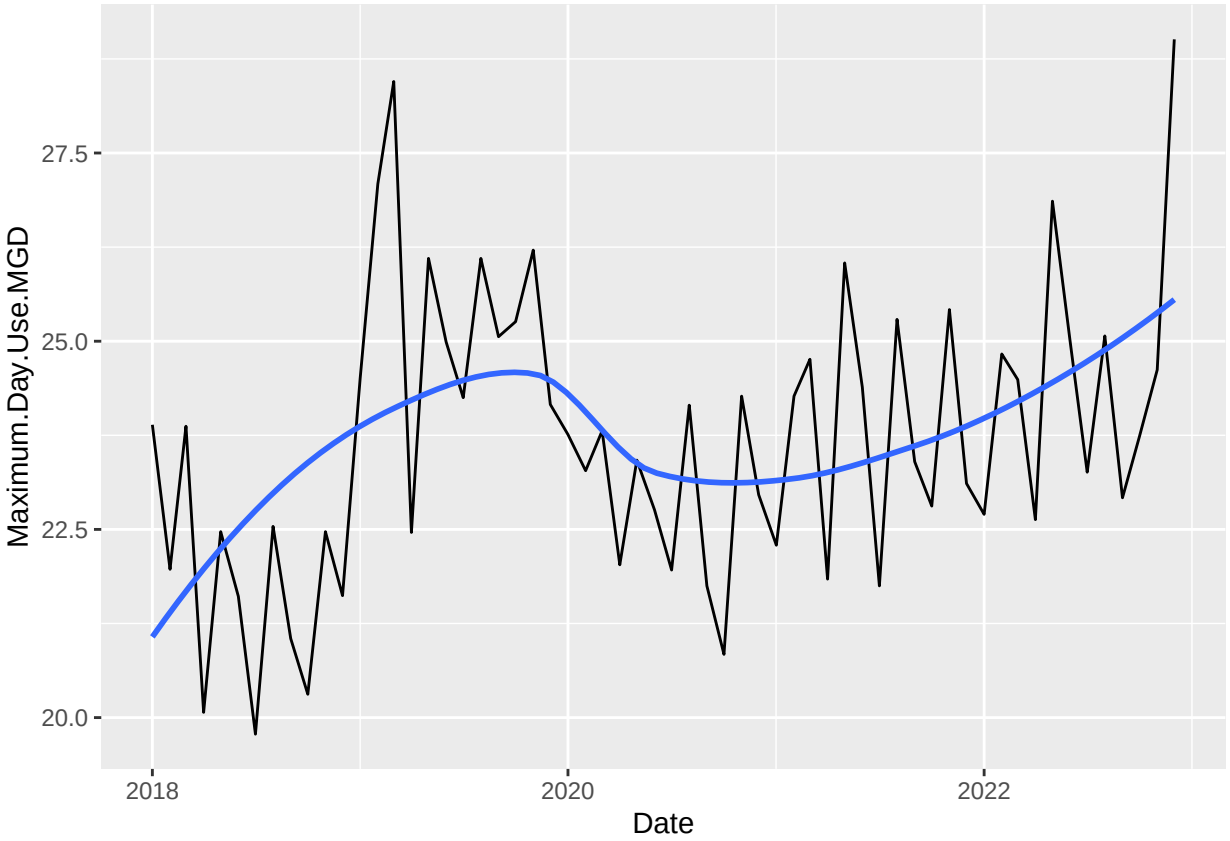



9. Use the code & function you created above to plot Asheville's max daily withdrawal by months for the years 2018 thru 2022.Add a smoothed line to the plot (method = 'loess').

   TIP: See Section 3.2 in the "10_Data_Scraping.Rmd" where we apply "map2()" to iteratively run a function over two inputs. Pipe the output of the map2() function to `bindrows()` to combine the dataframes into a single one.

```
#9
years <- rep(2018:2022)
PWSID <- rep.int('01-11-010',length(years))

MDW_Asheville_18thru22 <- map2(years, PWSID, scrape_ncwater) #I separated these two lines because when

MDW_Asheville_18thru22 <- bind_rows(MDW_Asheville_18thru22)

ggplot(MDW_Asheville_18thru22,
       aes(x = Date, y = Maximum.Day.Use.MGD)) +
  geom_line() +
   geom_smooth(method="loess",se=FALSE)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




Question: Just by looking at the plot (i.e. not running statistics), does Asheville have a trend in water usage over time? > Answer:It's hard to tell for this short of a time span, but it seems to be increasing gradually. >